

DISTRICT COMPREHENSIVE ASSESSMENT_ 2024/2025

MARKING GUIDE

Sector: ICT AND MULTIMEDIA

RQF Level: 5

Trade: SOFTWARE DEVELOPMENT

Marks:/ 100

Duration: 3 Hours

This exam paper is composed of twenty-four (24) questions distributed in three sections (A, B, C). Follow the instructions given below, and answer the indicated questions in each section for a total of 100 marks.

Section A: Seventeen (17) questions, all are compulsory .	55 marks
Section B: Five (5) questions, choose and answer any three (3).	30 marks
Section C: Two (2) questions, choose and answer any one (1).	15 marks

Allowed materials: Blue pen or black pen, Ruler, Non programmable Calculator, Pencil.

SECTION A: Attempt all questions

(55marks)

01 _____ is a field of AI that enables systems to learn from data and improve without explicit programming.

(3marks)

Answer: Machine learning

02 State Machine Learning life cycle stages

(5marks)

Answer:

- ✓ Problem Definition
- ✓ Data Collection
- ✓ Data Preprocessing
- ✓ Model Selection
- ✓ Model Training
- ✓ Model Evaluation
- ✓ Deployment
- ✓ Monitoring & Maintenance

03 List four applications of machine learning in real-world scenarios.

(4 marks)

Answer:

- ✓ Fraud detection in banking
- ✓ Personalized recommendations (Netflix, Amazon)
- ✓ Self-driving cars
- ✓ Medical diagnosis, etc

04 Name and explain the four types of machine learning. **(4marks)**

Answer:

- ✓ **Supervised Learning:** Uses labeled data (e.g., spam detection).
- ✓ **Unsupervised Learning:** No labels, finds patterns (e.g., clustering).
- ✓ **Semi-Supervised Learning:** Mix of labeled and unlabeled data.
- ✓ **Reinforcement Learning:** Learning through rewards (e.g., AlphaGo).

05 What is the purpose of a validation dataset in machine learning? **(3 marks)**

- a) Training the model
- b) Evaluating the final accuracy
- c) Fine-tuning hyperparameters
- d) Storing raw data

Answer: c) Fine-tuning hyperparameters

06 Which of the following is an example of supervised learning? **(3 marks)**

- a) K-Means Clustering
- b) Decision Trees
- c) Principal Component Analysis
- d) Hierarchical Clustering

Answer: b) Decision Trees

07 What does the term "overfitting" mean in machine learning? **(3marks)**

- a) The model performs well on test data but poorly on training data
- b) The model learns noise instead of the actual pattern
- c) The model performs equally well on both training and test sets
- d) The model is undertrained

Answer: b) The model learns noise instead of the actual pattern

08 In a classification task, which metric should you prioritize when dealing with an imbalanced dataset? **(3 marks)**

- a) Accuracy
- b) Precision & Recall
- c) Mean Squared Error
- d) R-squared Score

Answer: b) Precision & Recall

09 Which of the following is NOT a data visualization library in Python? **(3marks)**

- a) Matplotlib
- b) Seaborn
- c) NumPy
- d) Plotly

Answer: c) NumPy

- 10 What is the primary goal of Principal Component Analysis (PCA)? (3marks)
- a) Reduce overfitting
 - b) Reduce dimensionality while preserving variance
 - c) Increase model accuracy
 - d) Convert categorical data into numerical data

Answer: b) Reduce dimensionality while preserving variance

- 11 In reinforcement learning, what is the term for the entity that interacts with an environment to learn optimal actions? (3 marks)
- a) Actor
 - b) Agent
 - c) Trainer
 - d) Supervisor

Answer: b) Agent

- 12 Which of the following algorithms is used for classification problems? (3 marks)
- a) K-Means
 - b) Linear Regression
 - c) Logistic Regression
 - d) PCA

Answer: c) Logistic Regression

- 13 The six Vs of Big Data are Volume, Variety, Velocity, Veracity, _____, and _____. (3marks)

Answer: Value and Variability

- 14 A _____ function is used in neural networks to introduce non-linearity. (3 marks)

Answer: Activation

- 15 The purpose of hyperparameter tuning is to optimize the _____ of a machine learning model. (3 marks)

Answer: Performance

- 16 True/False: The accuracy metric is always the best measure of a model's performance. (3 marks)

Answer: False

- 17 True/False: Data normalization helps in improving model performance by scaling features to a common range. (3 marks)

Answer: True

Section B: Attempt any three (3) questions (30marks)

- 18 You trained a classification model and obtained the following confusion matrix: (10marks)

Actual\Predicted	Positive	Negative
Positive	90	10
Negative	20	80

Calculate the **accuracy, precision, and recall**.

Answer:

- **Accuracy** = $(90 + 80) / 200 = 85\%$
- **Precision** = $90 / (90 + 10) = 90\%$
- **Recall** = $90 / (90 + 20) = 81.8\%$

19 Write a Python program that:

(10marks)

- Loads the Breast Cancer dataset from sklearn.datasets.
- Splits it into training and testing data.
- Trains a Logistic Regression classifier.
- Prints the accuracy score.

Answer:

```
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

# Load dataset
data = load_breast_cancer()
X, y = data.data, data.target

# Split dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Train model
model = LogisticRegression(max_iter=10000)
model.fit(X_train, y_train)

# Predict and evaluate
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

20 Using Matplotlib, write Python code to generate a bar chart for the following data:

(10marks)

```
categories = ["Apples", "Bananas", "Cherries", "Dates"]
values = [30, 45, 20, 15]
```

Answer:

```
import matplotlib.pyplot as plt
plt.bar(categories, values)
plt.xlabel("Fruits")
plt.ylabel("Quantity")
plt.title("Fruit Inventory")
plt.show()
```

- 21 a) Given a dataset with missing values, describe how different types of visualizations can help identify data inconsistencies. **(10marks)**
b) Review a dataset containing sales records and determine the most effective visualization to detect anomalies in sales trends. Justify your choice.

a) Answer:

- **Heatmaps** can show missing values as gaps.
- **Box plots** can reveal outliers and inconsistencies.
- **Histograms** can display skewed distributions due to missing values.

b) Answer:

- Use a **time series line plot** to identify seasonal trends.
- Use a **box plot** to detect outliers.
- Use a **scatter plot** to find correlations between variables.

- 22 You are given a dataset containing sales data for different products across various regions. Use **Pandas** to visualize the total sales per region using a **bar chart**. **(10marks)**

Dataset (sales_data.csv):

Product	Region	Sales
A	North	1200
B	South	1500
C	East	900
D	West	1100
E	North	1300
F	South	1400
G	East	1000
H	West	1200

Answer:

```
import pandas as pd
import matplotlib.pyplot as plt
```

```
# Load the dataset
```

```
df = pd.read_csv("sales_data.csv")
```

```
# Aggregate total sales per region
```

```
sales_by_region = df.groupby("Region")["Sales"].sum()
```

```
# Plot a bar chart
```

```
plt.figure(figsize=(8, 5))
```

```
sales_by_region.plot(kind="bar", color=["blue", "green", "red", "purple"])
```

```
plt.title("Total Sales by Region")
```

```
plt.xlabel("Region")
```

```
plt.ylabel("Total Sales")
plt.xticks(rotation=0)
plt.grid(axis='y', linestyle='--', alpha=0.7)

# Show the plot
plt.show()
```

Section C: Attempt only one (1) question

(15marks)

23 Write a Python program that:

(15marks)

- Loads the **Iris dataset** using Scikit-learn.
- Splits it into training and testing data.
- Trains a **K-Nearest Neighbors (KNN) classifier**.
- Evaluates the accuracy of the model.

Answer:

```
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score

# Load dataset
iris = load_iris()
X, y = iris.data, iris.target

# Split dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Initialize and train model
model = KNeighborsClassifier(n_neighbors=3)
model.fit(X_train, y_train)

# Predict and evaluate
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print("Model Accuracy:", accuracy)
```

24 Write a Python script to perform **K-Means clustering** on a **synthetic dataset** using sklearn.datasets.make_blobs().

(15marks)

Answer:

```
from sklearn.cluster import KMeans
from sklearn.datasets import make_blobs
import matplotlib.pyplot as plt

# Generate dataset
X, _ = make_blobs(n_samples=200, centers=3, random_state=42)
```

```
# Apply K-Means
kmeans = KMeans(n_clusters=3, random_state=42)
y_pred = kmeans.fit_predict(X)

# Plot clusters
plt.scatter(X[:, 0], X[:, 1], c=y_pred, cmap='viridis')
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1],
c='red', marker='x', label='Centroids')
plt.legend()
plt.show()
```